



## How to Jump-Start Early Case Assessment: A Case Study based on TREC 2008

For over a decade, the National Institute of Standards and Technology ("NIST") has co-sponsored the Text Retrieval Conference ("TREC") to promote research in information retrieval from large text collections. Over the years, TREC, which itself grew out of an earlier NIST project called TIPSTER, has explored a number of specific text-based issues, such as identification of SPAM within large e-mail message populations and efficiently searching blog and world wide web content. Of greatest interest to the legal community, though, is the TREC Legal Track, which focuses on measuring the precision and accuracy of different methods used to retrieve relevant discovery documents in litigation matters.

Since 2006, the TREC Legal Track has conducted annual studies in which different search and analysis paradigms are applied to a large body of electronically searchable documents. The experiments are set up as mock lawsuits, and different teams are asked to research different issues in support of the larger case. Because the data set is standardized, results of the work done by each team can be compared against neutral results to compare the efficacy of each approach.

As part of the studies, the TREC Legal Track invites technology solutions providers to enter teams using their products. The work of these teams enables the comparison of different approaches and technologies, including traditional keyword search.

Studies of keyword-based retrieval have demonstrated the inability of keyword search to find most of the relevant documents, while retrieving a high proportion of documents that are not relevant (see for example the famous Blair & Maron study from 1985 and [The Sedona Conference® Best Practices Commentary on Search & Retrieval Methods](#)). The TREC Legal Track explores the ability of new technologies to provide greater consistency and accuracy in these types of projects.

In widely quoted comments from *Victor Stanley v. Creative Pipe* relating to the discovery of electronically stored information, Judge Grimm noted: “[T]here is room for optimism that as search and information retrieval methodologies are studied and tested, this will result in identifying those that are most effective and least expensive to employ for a variety of ESI discovery tasks. Such a study has been underway since 2006...This project, known as the Text Retrieval Conference (TREC) ... Legal Track, [is] a research effort aimed at studying the e-discovery review process to evaluate the effectiveness of a wide array of search methodologies ... This project can be expected to identify both cost effective and reliable search and information retrieval methodologies and best practice recommendations, which, if adhered to, certainly would support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes.”

Equivio’s new Equivio>Relevance analytical tool was completed too late to enter into the 2008 TREC Legal Track. However, Equivio recently applied the tool to the same data set and to one of the standardized topics that were used by last summer’s teams. The results demonstrate the dramatic superiority of machine learning techniques over traditional manual keyword search approaches to early case assessment.

## **The Trial**

The 2008 TREC Legal Track invited teams to find all relevant documents relating to one of several topics. Four teams, two using technology-based analytical engines and two using only traditional human reviewers, worked with Topic 103, “All documents which refer to ‘in-store’, ‘on-counter’, ‘point-of-sale’, or other retail marketing campaigns for cigarettes.” This is the topic that Equivio used for its trial.

Equivio’s first step was to import the 2008 TREC Legal Track document collection into Equivio>Relevance. These 6,910,192 documents comprise the IIT Complex Document Information Processing (CDIP) Test Collection, version 1.0. This collection is based on documents released under the tobacco “Master Settlement Agreement.” The same document set had been used previously in the 2006 and 2007 TREC Legal Track experiments and represents one of the few large non-proprietary document collections available for evaluations of text retrieval and analysis solutions. However, because of the age of document collection and the way in which these materials were processed at the time of their production in litigation, these documents are not native ESI, but rather are OCR text in XML format (non-searchable digitized images of the hardcopy originals are not used by TREC). The OCR errors impose an additional challenge for document

retrieval technologies. Furthermore, metadata is generally not available. While the documents have elements rendered in XML, they are not true native files. As such, the documents contain little if any meaningful fielded metadata, and their relevance must be ascertained from the body of the documents.

During the Summer 2008 TREC Legal Track, each participating team had a bank of up to ten (10) hours of consulting time with a human subject matter expert designated the "Topic Authority" whom they could use to help fine-tune the team's understanding of the topic for which they needed to find relevant documents. By conducting its analysis after the close of the 2008 TREC study, the Equivio team had no Topic Authority for assistance in fine-tuning the initial concepts and search terms that might define relevance.

The Equivio>Relevance product is an expert-guided system to assess document relevance. The software feeds statistically selected samples of documents to an expert, typically an attorney familiar with the case. The expert designates each of the sample documents as relevant or not. The samples are used to "train" the software as to the characteristics of relevant documents. At the end of the process, the software produces a relevance score for each document in the collection.

Equivio selected a seasoned litigation attorney as the "expert". In order to learn about the case, the attorney used the mock briefs that had been prepared by the TREC organizers and the case workup that had been provided to all teams working on Topic 103.

Equivio>Relevance fed the attorney samples of documents from the 6.9 million document collection. These samples were broken into 35 batches, each containing 40 documents. While 1,400 documents may sound like a lot of material to review, the attorney made only a simple relevance determination based on the Topic 103 parameters, with no sub-issues or other distractions. Importantly, the sample represented a very small proportion of the seven million of documents in the collection. In the end, the attorney's relevance analysis took about 18 hours, spread over two 9-hour days, to complete the sample document designations which, in turn, were used to "train" the Equivio>Relevance tool regarding the factors that contribute to document relevance in Topic 103.

After the Equivio>Relevance training process was completed, Equivio>Relevance began its analysis of the complete 2008 TREC Legal Track document collection. Processing took about 10.5 hours of machine time, after which the results were available for human review.

## Trial Results

During the 2008 TREC Legal Track, team results were measured by three basic criteria: (1) recall; (2) precision; and (3) "F-measure". Recall measures how many of the total relevant documents within the collection have been retrieved and is expressed as a percentage (i.e., a 20% recall would mean that a team found 20% of all relevant documents within the collection). Precision, which is also expressed as a percentage, measures the accuracy of the documents categorized by the team as relevant to the topic they were researching (i.e., a 30% precision rate would mean that 30% of the documents flagged as relevant were in fact relevant). There is an inherent trade-off between recall and precision – as recall increases, precision decreases, and vice-versa. The F-measure is a mathematically calculated value balancing (not averaging) the recall and precision rates.

Because of the millions of documents in the CDIP Test Collection, not every document can be exhaustively categorized in advance of the exercises. Instead, a panel of human "Oracles" analyzed 6,068 of the documents. Recall and precision rates were calculated by measuring team results against those of the Oracle. While a team could appeal a difference of opinion they had regarding categorization of a specific document, the Oracle's decision after further review of the document was final and binding.

Equivio>Relevance, even without consulting assistance from the Topic Authority, achieved outstanding results in all three categories: recall, precision, and F-measure.

	<b>Recall</b>	<b>Precision</b>	<b>F-measure</b>
<b>Equivio&gt;Relevance</b>	<b>71.0%</b>	<b>81.4%</b>	<b>75.8%</b>

Full details of the TREC 2008 results can be found in the TREC document: Overview of the TREC 2008 Legal Track (see Table 15 on page 10 of the report, which is available at <http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf>)

Additional analysis within the 2008 TREC Legal Track results also highlights the dramatic difference between traditional keyword search methodology and the Equivio>Relevance-based approach. Keyword search generated an average recall rate of approximately 24% and an average precision rate of 28% (see Table 1 in the Overview document). These numbers indicate the risk of standard keyword-based retrieval – that is, more than 75% of the relevant documents are lost in the process, and are not subject to detailed review. Similarly, given the low precision rates of keyword searches, over 70% of the documents reviewed are not in fact relevant. Given the existence of alternative technologies, the low precision rates associated with keyword techniques are imposing an unnecessarily burdensome review cost on litigants. By comparison, Equivio>Relevance attained almost triple the recall and precision rates of keyword search.

## Conclusion

The TREC initiative is an important component in the industry's ongoing effort to assess and measure the potential contribution of new approaches and technologies that can help reduce the cost and risk of litigation. Equivio plans to participate in the TREC 2009 Legal Track study. The TREC 2009 competition will take place over the summer. The teams are required to post their submissions in September, with final results due to be published in February 2010.

## About Equivio

Equivio enables the management of data redundancy in content-centric business processes. Equivio's technology zooms in on unique data, allowing you to read less, think more, win big™.

With products for grouping near-duplicates, capturing email threads, determining document relevance and automating prioritization, Equivio powers a broad range of business applications, including e-discovery, records management, email archiving, data retention and intelligence.

To learn more about winning with Equivio, email [info@equivio.com](mailto:info@equivio.com) or visit [www.equivio.com](http://www.equivio.com).

Equivio Inc.  
5260G Nicholson Lane, Suite 150  
Kensington, MD 20895  
USA